(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 020 847 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
19.07.2000 Bulletin 2000/29

(51) Int Cl.7: **G10L 15/22**, G10L 15/08

(21) Application number: **00660008.4**

(22) Date of filing: **18.01.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **18.01.1999 FI 990077**

(71) Applicant: **NOKIA MOBILE PHONES LTD.**
02150 Espoo (FI)

(72) Inventors:
- **Laurila, Kari**
  33720 Tampere (FI)
- **Iso-Sipilä, Juha**
  33180 Tampere (FI)

(74) Representative: **Pursiainen, Timo Pekka**
Tampereen Patenttitoimisto Oy,
Hermiankatu 6
33720 Tampere (FI)

(54) Method for multistage speech recognition using confidence measures

(57) In a method for recognizing speech commands, in which a group of command words selectable by speech commands are defined, a time window is defined, within which the recognition of the speech command is performed. In a first recognition stage in the method, the recognition result of the first recognition stage is selected, for which a first confidence value is determined. Further in the method, a first threshold value (Y) is determined, with which said first confidence value is compared. If said first confidence value is greater than or equal to said first threshold value (Y), the recognition result of the first recognition stage is selected as the recognition result of the speech command. If said first confidence value is smaller than said first threshold value (Y), a second recognition stage is performed for the speech command, wherein said time window is extended, and a recognition result is selected for the second recognition stage. A second confidence value is determined for the recognition result of the second recognition stage and compared with said threshold value (Y). If said second confidence value is greater than or equal to said first threshold value (Y), the command word selected at the second stage is selected as the recognition result for the speech command. If said second confidence value is smaller than said first threshold value (Y), a comparison stage is performed, wherein the first and second recognition results are compared to find out at which probability they are substantially the same, wherein if the probability exceeds a predetermined value, the command word selected at the second stage is selected as the recognition result for the speech command.

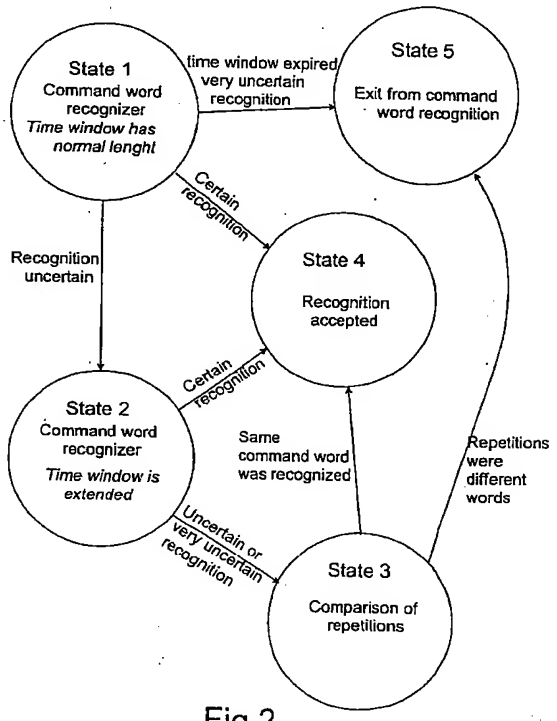

Fig 2

EP 1 020 847 A2

0.9, the recognition is not accepted and the user may have to utter the command several times until the recognition probability threshold is exceeded and the speech recognition device accepts the command, even though the probability may have been very close to the acceptable value. This is very disturbing to the user.

[0007] Furthermore, the speech recognition is hampered by the fact that different users utter the same words in different ways, wherein the speech recognition device works better when used by one user than when used by another user. In practice, it is very difficult with the presently known techniques to adjust the certainty levels of speech recognition devices to consider all users. When adjusting the required certainty level *e.g.* for the word "yes" in speech recognition devices of prior art, the required threshold is typically set according to so-called worst speakers. Thus, the problem emerges that words close to the word "yes" also become incorrectly accepted. The problem is aggravated by the fact that in some situations, mere background noise may be recognized as command words. In speech recognition devices of prior art, the aim is to find a suitable balance in which a certain part of the users have great problems in having their words accepted and the number of incorrectly accepted words is sufficiently small. If the speech recognition device is adjusted in a way that a minimum number of users have problems in having their words accepted, this means in practice that the number of incorrectly accepted words will increase. Correspondingly, if the aim is set at as faultless a recognition as possible, an increasing number of users will have difficulties in having commands uttered by them accepted.

[0008] In speech recognition, errors are generally classified in three categories:

- Insertion Error
  The user says nothing but a command word is recognized in spite of this, or the user says a word which is not a command word and still a command word is recognized.

- Deletion Error
  The user says a command word but nothing is recognized.

- Substitution Error
  The command word uttered by the user is recognized as another command word.

[0009] In a theoretical optimum solution, the speech recognition device makes none of the above-mentioned errors. However, in practical situations, as was already presented above, the speech recognition device makes errors of all the said types. For usability of the user interface, it is important to design the speech recognition device in a way that the relative shares of the different error types are optimal. For example in speech activation, where a speech-activated device waits even for hours for a certain activation word, it is important that the device is not erroneously activated at random. Furthermore, it is important that the command words uttered by the user are recognized at good accuracy. In this case, however, it is more important that no erroneous activations take place. In practice, this means that the user must repeat the uttered command word more often so that it would be recognized correctly at a sufficient probability.

[0010] In the recognition of a numerical sequence, almost all errors are equally significant. Any error in the recognition of the numbers in a sequence results in a false numerical sequence. Also the situation that the user says nothing and still a number is recognized, is inconvenient for the user. However, a situation in which the user utters a number indistinctly and the number is not recognized, can be corrected by the user by uttering the numbers more distinctly.

[0011] The recognition of a single command word is presently a very typical function implemented by speech recognition. For example, the speech recognition device may ask the user: "Do you want to receive a call?", to which the user is expected to reply either "yes" or "no". In such situations where there are very few alternative command words, the command words are often recognized correctly, if at all. In other words, the number of substitution errors in such a situation is very small. The greatest problem in the recognition of single command words is that an uttered command is not recognized at all, or an irrelevant word is recognized as a command word. In the following, there are three different alternative situations of this example:

1) A speech-controlled device asks the user: "Do you want to receive a call?", to which the user replies indistinctly: "Yes ... ye-". The device does not recognize the user's reply and asks the user again: "Do you want to receive a call? Say yes or no." Thus the user may be easily frustrated, if the device often asks the user to repeat the command word uttered.

2) The device asks the user again: "Do you want to receive a call?", to which the user responds distinctly "yes". However, the device did not recognize this for certain and wants a confirmation: "Did you say yes?", to which the user replies again "yes". Even now, no reliable recognition was made, so the device asks again: "Did you say yes?". The user must repeat again the reply "yes", for the device to complete the recognition.

3) Still in a third example situation, the speech-controlled device asks the user, if s/he wants to receive a call. To this, the user mumbles something vague, and in spite of this, the device interprets the user's utterance as the command word "yes" and informs the user "All right, the call will be connected". Thus, in this situation, the interpretation of the device of

was recognized. Figure 1 illustrates the determination of the confidence of this recognition by using threshold values and said confidence value. In the method according to an advantageous embodiment of the invention, a first threshold value is determined, indicated in the appended figure with the reference Y. It is the limit determined for the confidence value that the recognition is positive (the confidence value is greater than or equal to the first threshold value Y). In the method according to a second advantageous embodiment of the invention, also a second threshold value is determined, indicated in the appended figure with the reference A. This indicates whether the recognition was uncertain (the confidence value is greater than or equal to the second threshold value A but smaller than the first threshold value Y) or very uncertain (the threshold value is smaller than the second threshold value A).

[0017] In the state machine presentation of Fig. 2, state 1 represents the recognition of a command word. At this stage of recognizing the command word, probabilities are determined on the basis of the command word uttered by the user for different command words in the vocabulary of the speech recognition device. As the command word corresponding to the speech command uttered by the user is selected, preliminarily, the command word with the greatest probability. For the selected command word, said confidence value is determined and compared with the first threshold value Y and the second threshold value A to deduce whether the recognition was certain, uncertain or very uncertain. If the confidence value is greater than or equal to the first threshold value 6, the operation moves on to state 4 to accept the recognition. If the confidence value remained smaller than the second threshold value A, the operation moves on to state 5 to exit from the command word recognition, *i.e.* to reject the recognition. If the confidence value was greater than or equal to the second threshold value A but smaller than the first threshold value Y, the recognition was uncertain, and the operation moves on to state 2. Thus, the time window is extended, *i.e.* the user will have more time to say the uttered command word again. The operation can move on to this state 2 also because of an incorrect word, *e.g.* as a result of a word uttered very unclearly by the user or as a result of incorrect recognition caused by background noise. In this state 2, the repetition of the command word is waited for the time of the extended time window. If the user uttered a command word again in this time window, the command word is recognized and the confidence value is calculated, as presented above in connection with the state 1. If at this stage the calculated confidence value indicates that the command word uttered at this second stage is recognized with sufficient confidence, the operation moves on to the state 4 and the recognition is accepted. For example, in a situation when the user may have said something vague in the state 1 but has uttered the correct command word clearly in the state 2, the recognition can be made solely on the basis of this command word uttered in the state 2. Thus, no comparison will be made between the first and second command words uttered, because this would easily lead to a less secure recognition decision.

[0018] Nevertheless, if the command word cannot be recognized with sufficient confidence in the state 2, the operation moves on to state 3 for comparison of the repeated command words. If this comparison indicates that the command word repeated by the user was very close to the command word first said by the user, *i.e.* the same word was probably uttered twice in succession, the recognition is accepted and the operation moves on to the state 4. However, if the comparison indicates that the user has probably not said the same word twice, the operation moves on to the state 5 and the recognition is rejected.

[0019] Consequently, in the method of the invention, when the first step indicates an uncertain recognition, a second recognition is made preferably by a recognition method known as such. If this second step provides no sufficient certainty of the recognition, a comparison of the repetitions is made advantageously in the following way. In the state 1, feature vectors formed of the command word uttered by the user are stored in a speech response memory 4 (Fig. 5). Such feature vectors are distinguished from the speech typically at intervals of ca. 10 ms, *i.e.* ca. 100 feature vectors per second. Also in the state 2, feature vectors formed of the command word uttered at this stage are stored in the speech response memory 4. After this, the recognition moves on to the state 3, in which these feature vectors stored in the memory are compared preferably by dynamic time warping. Figure 3 illustrates this dynamic time warping of feature vectors in a reduced manner. At the top of the figure are shown feature vectors produced by the first recognition, indicated with the reference number V1, and correspondingly, at the bottom of the figure are shown feature vectors produced by the second recognition and indicated with the reference number V2. In this example, the first word was longer than the second word, *i.e.* the user has said the word faster at the second stage, or the words involved are different. Thus, for the feature vectors of the shorter word, in this case the second word, one or more corresponding feature vectors are found from the longer word by time warping the feature vectors of the two words in a way that they correspond to each other optimally. In this example, these time warping results are indicated by broken lines in Fig. 3. The distance between the words is calculated *e.g.* as a Euclidean distance between the warped feature vectors. If the calculated distance is small, it can be assumed that the words in question are different. Figure 4 shows an example of this comparison as a histogram. The histogram includes two different comparisons: the comparison between two identical words (shown in solid lines) and a comparison between two different words (shown in broken lines). The horizontal axis is the logarithmic value of the calculated distance between the fea-

ognized with sufficient confidence and the user does not repeat the command word within the time limit, the command word is not accepted. In this case too no control signal is transmitted to the control block 16.

[0025] To increase the convenience of use of the wireless communication device 1 in those cases where the first recognition of the command word did not provide a sufficiently reliable recognition, the user can be informed of the failure of the recognition of the first stage and be requested to utter the command word again. The wireless communication device 1 forms *e.g.* an audio message with a speech synthesizer 8 and/or a visual message on a display means 13. The wireless communication device 1 can inform the user with an audio and/or visual signal also in a situation where the recognition was successful. Thus it will not remain obscure to the user whether the recognition was successful or not. This is particularly useful under noisy use conditions.

[0026] Warping of feature vectors and calculating of distances between words is prior art known *per se,* why these are not disclosed here in more detail. It is obvious that the present invention is not limited solely to the embodiments presented above but it can be modified within the scope of the appended claims.

Claims

1. A method for recognizing speech commands by using a time window, which is extendable when needed, in which method a group of command words selectable by speech commands are defined, a time window is defined, within which the recognition of the speech command is performed, and a first recognition stage is performed, in which the recognition result of the first recognition stage is selected, **characterized** in that further in the method:

   a) a first confidence value is determined for the recognition result of the first recognition stage,
   b) a first threshold value (Y) is determined,
   c) said first confidence value is compared with said first threshold value (Y),
   d) if said first confidence value is greater than or equal to said first threshold value (Y), the recognition result of the first recognition stage is selected as the recognition result of the speech command,
   e) if said first confidence value is smaller than said first threshold value (Y), a second recognition stage is performed for the speech command, wherein
   f) said time window is extended, and
   g) a second confidence value is determined for the recognition result of the second recognition stage,
   i) said second confidence value is compared with said threshold value (Y),
   j) if said second confidence value is greater than or equal to said first threshold value (Y), the command word selected at the second stage is selected as the recognition result for the speech command,
   k) if said second confidence value is smaller than said first threshold value (Y), a comparison stage is performed, wherein
   l) the first and second recognition results are compared to find out at which probability they are substantially the same, wherein if the probability exceeds a predetermined value, the command word selected at the second stage is selected as the recognition result for the speech command.

2. The method according to claim 1, **characterized** in that at said recognition stages, a probability is determined for one or several said command words, at which the speech command uttered by the user corresponds to said command word, wherein the command word with the greatest determined probability is selected as the recognition result of said recognition stages.

3. The method according to claim 1 or 2, **characterized** in that in the method an additional second threshold value (A) is determined, wherein the stages e) to k) are performed only if said first confidence value is greater than said second threshold value (A).

4. The method according to claim 3, **characterized** in that the comparison stage k) is performed only if said second confidence value is greater than said second threshold value (A).

5. The method according to any of the claims 1 to 4, **characterized** in that for determining the first confidence value, a probability is determined for the first speech command being background noise, wherein the first confidence value is formed on the basis of the probability determined for the command word selected as the recognition result of the first recognition stage, and the background noise probability.

6. The method according to any of the claims 1 to 5, **characterized** in that for determining the second confidence value, a probability is determined for the second speech command being background noise, wherein the second confidence value is formed on the basis of the probability determined for the command word selected as the recognition result of the second recognition stage, and the background noise probability.

7. A speech recognition device, in which a vocabulary of selectable command words is defined, the device

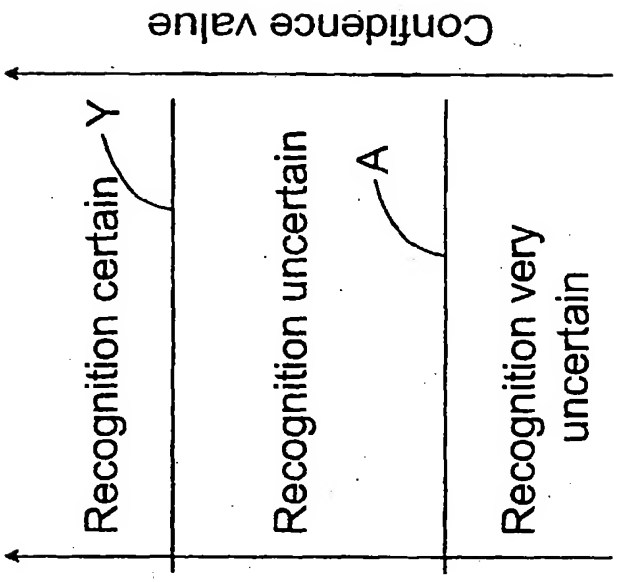

Fig 1

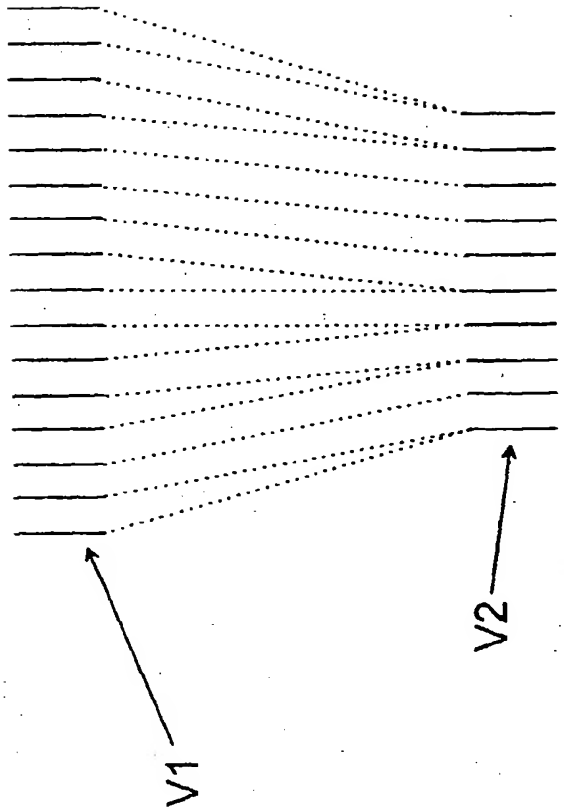

Fig 3

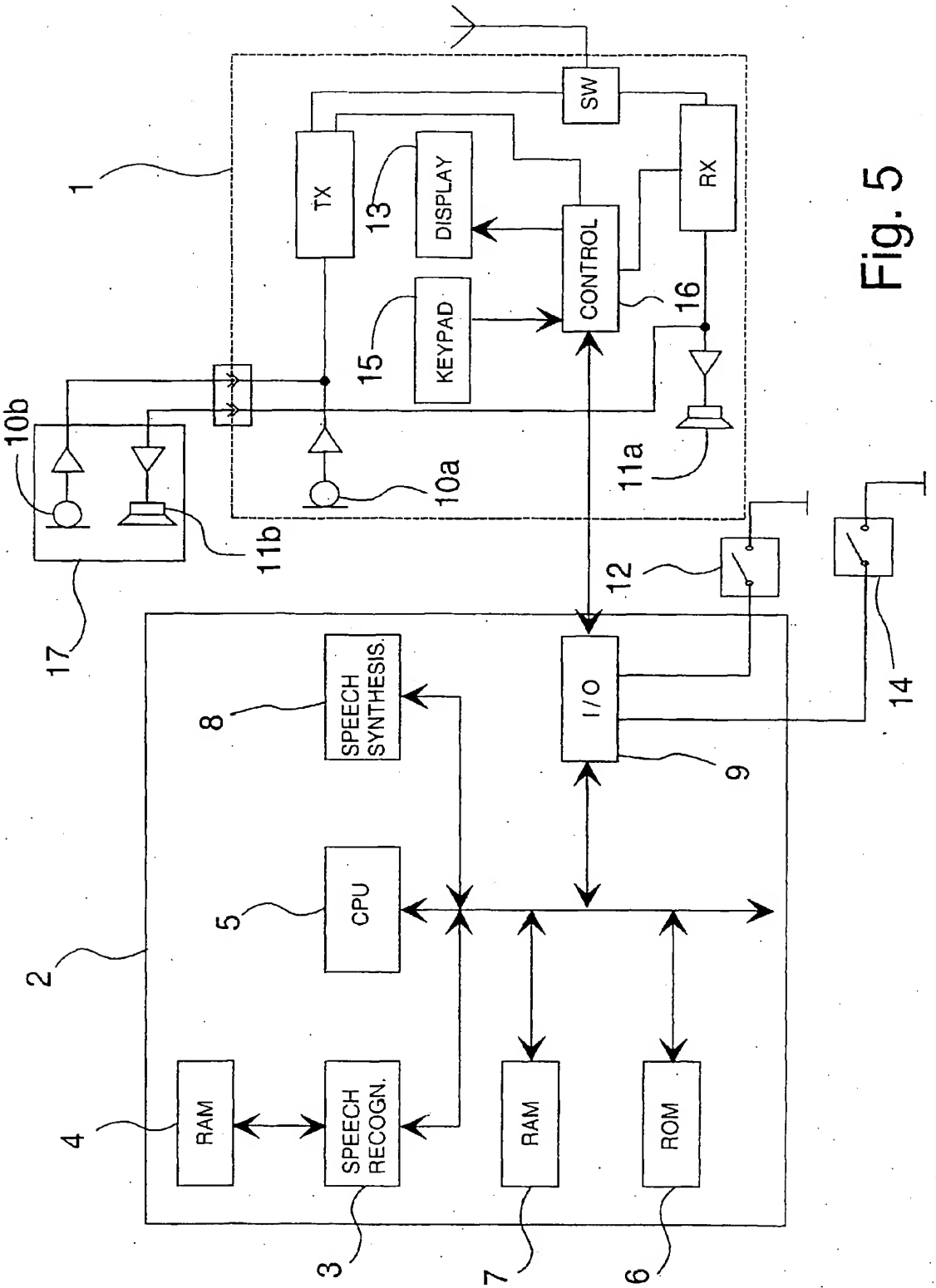

Fig. 5

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 020 847 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
02.05.2001 Bulletin 2001/18

(51) Int Cl.7: **G10L 15/22**, G10L 15/08

(43) Date of publication A2:
19.07.2000 Bulletin 2000/29

(21) Application number: **00660008.4**

(22) Date of filing: **18.01.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **18.01.1999 FI 990077**

(71) Applicant: **NOKIA MOBILE PHONES LTD.**
02150 Espoo (FI)

(72) Inventors:
- **Laurila, Kari**
  33720 Tampere (FI)
- **Iso-Sipilä, Juha**
  33180 Tampere (FI)

(74) Representative: **Purslainen, Timo Pekka**
Tampereen Patenttitoimisto Oy,
Hermiankatu 6
33720 Tampere (FI)

(54) Method for multistage speech recognition using confidence measures

(57) In a speech recognition system the recognition hypothesis extracted using a first time window is used to calculate a first confidence measure. If this confidence is low, a second recognition stage with an extended time window is applied to the group of words. If the confidence of the second stage hypothesis is again low, a comparison is made to find out if the first and second hypothesis are substantially the same. If not the recognizer outputs the second stage hypothesis.

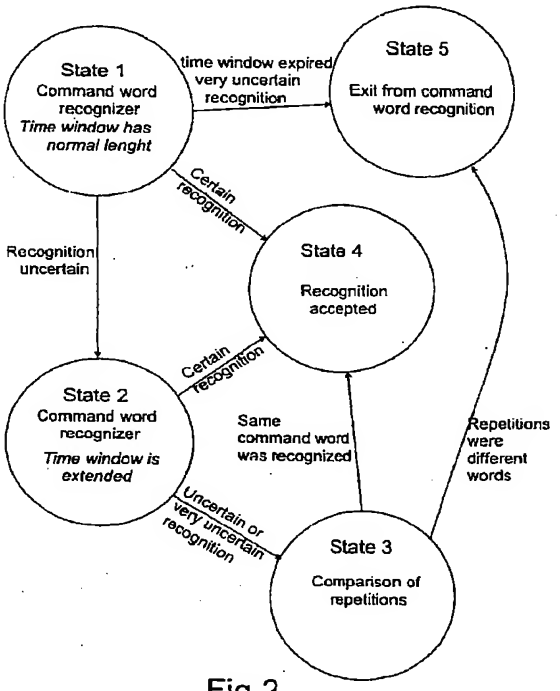

Fig 2

Printed by Jouve, 75001 PARIS (FR)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 00 66 0008

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

06-03-2001

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0651372 A	03-05-1995	US 5566272 A	15-10-1996
		CA 2131600 A,C	28-04-1995
		CN 1115902 A	31-01-1996
		JP 7181994 A	21-07-1995
US 5737724 A	07-04-1998	CA 2117932 A,C	25-05-1995
		EP 0655732 A	31-05-1995
		JP 7199985 A	04-08-1995
US 5794194 A	11-08-1998	JP 3004023 B	31-01-2000
		JP 3167600 A	19-07-1991
		DE 69026474 D	15-05-1996
		DE 69026474 T	19-09-1996
		EP 0430615 A	05-06-1991
US 6122613 A	19-09-2000	EP 0954848 A	10-11-1999
		WO 9834217 A	06-08-1998

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82